Video-Data Knowledge Modelling & Discovery

J.L. Patino*, H. Benhadda†, E. Corvee*, F. Bremond*, M. Thonnat*

*INRIA, FRANCE {jlpatino, Etienne.Corvee, Francois.Bremond, Monique.Thonnat}@sophia.inria.fr
†Thales Comunication, FRANCE Hamid.Benhadda@fr.thalesgroup.com

Keywords: Tracking, Information, Representation, Behaviour, Clustering.

Abstract

Most video applications fail to capture in an efficient knowledge representation model interactions between subjects themselves and interactions between subjects and contextual objects of the observed scene. In this paper we propose a knowledge modelling format which allows efficient knowledge representation. Furthermore, we show how advanced algorithms of knowledge discovery can be applied following the proposed format.

1 Introduction

A challenging problem in the management of large video collections is the ability to automatically extract, model and store structured knowledge from the video streams in a meaningful way. Vast energies have been expended in the last years in the development of video analysis and video databases; however, the utility to end-users is still limited because such systems mostly index video using low-level features which limit the potential information to the end-user. Few systems track moving objects in the scene and index the observed objects together with their position over time [1, 5, 6, 9, 15, 17]. This gives indeed a spatiotemporal representtation of the video content. However some information is lost because the interaction between moving objects and their environment has been only partially studied and is rarely used as a feature to describe video content. Li et *al.* [11], for instance, studied the interaction between mobile objects modelling the history of an object but no further analysis is done on the contextual objects of the scene. Lin et *al.* [12] give an interesting knowledge representation of the video by describing separately the video scene and the moving objects but again the interaction is not studied. Liu et *al.* [13] also proposes a structured representation of a moving object as a tuple with six features. Behaviour discovery can be achieved but not about the interaction between moving objects or with contextual objects. If both kinds of interaction are modelled and represented in a proper way, a higher level of semantic content in the video can be presented to the end-user.

In this paper we propose a knowledge modelling format which allows efficient knowledge representation. In our approach, a first layer of knowledge can be extracted directly on-line from the raw data streams. A second layer of higher semantic knowledge is defined from longer off-line analysis and set in the proposed format. Namely, we divide all information into three tables: Mobile objects, Contextual objects and Events. This makes indeed a major difference with previous video interpretation systems such as PRISMATICA [18], VISOR-BASE [16] and our own previous system ADVISOR [7]. In such systems, the efforts were concentrated on efficient on-line detection of a series of events such as overcrowding/congestion; unusual or forbidden directions of motion; stationarity of people; fighting between persons; vandalism,... but monitoring the interaction between people and contextual objects of the scene and the evolution of use of these contextual objects was not addressed, which we achieve off-line thanks to the proposed representation format. Furthermore, we show how advanced algorithms of knowledge discovery can be applied following the proposed format to find out complex events difficult to see at first sight from the low-level features.

This research has been done in the framework of the CARETAKER project, which is an European initiative to provide an efficient tool for the management of large multimedia collections. Such system could be used in applications such as surveillance and safety issues, in urban/environment planning, resource optimization, disabled/elderly person monitoring. Currently it is being tested on large underground video recordings (GTT metro, Torino, Italy and ATAC metro, Roma, Italy).

The rest of the paper is structured in the following way.
In section 2 we present the overall architecture of the proposed approach. While the on-line analysis is explained in section 3, the off-line counterpart is detailed in section 4. Results on annotated and real data are presented in section 5. The proposed method is discussed in section 6 and our final conclusions are also given.

2 General structure of the proposed approach

There are three main components which define our approach: The data acquisition; the on-line analysis of video streams; the long-term off-line analysis. The graphical schema is shown in Figure 1. Video streams are directly fed into our on-line analysis system for real time detection of objects and events in the scene. This procedure goes on a frame-by-frame

basis and the results are stored into a specific on-line database. At this level, detected events already contain semantic information describing the interaction between objects and the contextual information of the scene. This is the first layer of semantic information in our system. The long-term analysis of detected objects and events retrieved from the on-line database will deliver new information difficult to see directly on the video streams. This constitutes a second layer of semantic information. Statistical measures such as most frequent events, time spent by users to interact with contextual objects of the scene are measured. Also the trajectories undertaken by the users are characterised. All this information is set up in a suitable knowledge representation model from which complex relationships can be discovered using relational analysis.

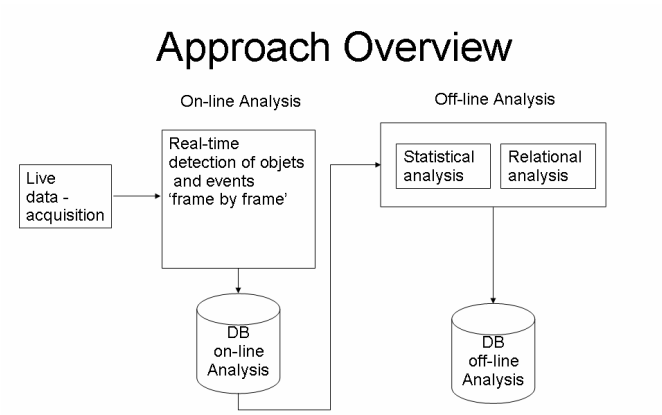


Figure 1: Architecture of the proposed approach.

3 Real-time Object/Event detection

3.1 Multiple objects tracking

Tracking several mobile objects evolving in a scene is a difficult task to perform. Motion detectors often fails in detecting accurately moving objects referred to as 'mobiles' which induces mistracks of the mobiles. Such errors can be caused by shadows or more importantly by static (when a mobile object is hidden by a background object) or by dynamic (when several mobiles projections onto the image plane overlap) occlusion [8].

The tracking algorithm builds a temporal graph of connected objects over time to cope with the problems encountered during tracking. The detected objects are connected between each pair of successive frames by a frame to frame (F2F) tracker [2]. The links between objects are associated with a weight (i.e. a matching likelihood) computed from three criteria: the similitude between their semantic classes, 2D dimension differences and 3D distance difference on the ground plane.

The graph of linked objects is analysed by the tracking algorithm also referred to as the Long Term Tracker which builds paths of each mobiles according to the links established by the F2F tracker. The best path is then taken out as the trajectory of the related mobiles. Figure 2 shows a tracked person labelled 1 and a tracked crowd of people labelled 534. Figure 2 shows also a tracked person, labelled 24 with two new objects: a group of persons labelled 58 and an unclassified tracked object labelled 68. Due to the poor contrasted lower part of the group of persons, this group has been segmented into two tracked objects (instead of one labelled 24 and 68).

3.2 Event detection

In this application, 10 'gates' (i.e. the access to the platform), 2 ticket vending machines and one platform (or central hall) compose the scene. The platform delimits the ground floor where all mobiles are allowed to evolve. The detected events are the following:

- 'inside_zone(o, z)': when an object 'o' is in the zone 'z'.
-'stays_inside_zone(o,z,T1)': when the event 'inside_ zone(o,z)' is being detected successively for at least T1 seconds
- 'close_to(o, eq, D)': when the 3D distance of an object location on the ground plane is less than the maximum distance allowed, D, from an equipment object 'eq'
- 'stays_at': when the event 'close(o, eq, Dmax, T2)' is being consecutively detected for at least T2 seconds.
- 'crowding_in_zone': when the event 'stays_inside_zone (crowd, z, T3)' is detected for at least T3 seconds.

Where,

* object o={p, g, c, l, t, u} with p=person, g=group, c=crowd l=luggage, t=train, and u=unknown.
* zone z={platform, validating_zone, vending_zone}
* equipment eq={g1, …, g10, vm1, vm2 } where 'gi' is the ith gate and vmi is the ith vending machine.

T1=60 s, D=1m50, T2=5 s, T3=120 s

Figure 2: A person and a crowd are tracked. The event 'person' stays at gates has been detected.

Figure 3: The event 'group stays at vending machine' has been detected.

Figure 2 shows the tracked person labelled 1 which remained long enough in front of the validating ticket machines labelled 'Gates' so that the event 'stays_at(p, gates, 5 seconds)' is detected. The tracked group of persons labelled 24 in Figure 3 is interacting with the vending machine number 2 long enough for the event 'stays_at(g, vm2, 5 seconds)' to be detected. In both these figures, the primitive event 'inside_zone' was not shown but was also detected for the remaining objects present in the hall.

4 Knowledge discovery

The second layer of analysis in our approach is related to the knowledge discovery of higher semantic events from off-line analysis of activity recorded over periods of time that can span for instance from some minutes to a whole day. The knowledge representation format we propose, partially answers this question by giving a statistical overview of the activities in the scene. For the analysis of more complex relationships between the objects observed in the scene we employ the relational analysis clustering technique.

4.1 Knowledge representation format

There are two main types of concepts to be represented from the video: physical objects of the observed scene and video events occurring in the scene. The former category can still be further subdivided into two types of physical objects of interest: mobile and contextual objects. Mobile objects of interest are the source of action occurring in the scene. Contextual objects are parts of the empty scene model corresponding to the static environment of the scene. Thus, for the off-line analysis of both types of concepts, with the aim of setting the data in a suitable format to achieve Knowledge Discovery, we separate the information corresponding to the activities occurring over a period of time on three different semantic tables, namely mobile objects, contextual objects and video events.

4.1.1 Mobile objects

A mobile object can be represented as an eight-tuple(m_id, m_type, m_start, m_end, m_shape, m_involved_events_id, m_significant_event, m_trajectory)
where
m_id. The identifier label for the object.
m_type. The class the object belongs to: Person, Group, Crowd or Luggage.
m_start. Time the object is first seen.
m_end. Time the object is last seen.
m_shape. The label describing the object's shape depending on the object's ratio height/width.
m_involved_events_id. All occurring Events related to the identified object.
m_significant_event. The most significant event among all events. This is calculated as the most frequent event related to the mobile object.
m_trajectory_type. The trajectory pattern characterising the object. For this purpose we have applied a hierarchical clustering algorithm to find different patterns of trajectories and thus have a comprehensive, compact, and flexible representation suitable also for further analysis as opposed to many video systems which actually store the sequence of object locations for each frame of the video, which is a cumbersome representation with no semantic information.
If the dataset is made up of m objects, the trajectory for object *i* in this dataset is defined as the set of points $[x_i(t),y_i(t)]$; x and y are time series vectors whose length is not equal for all objects as the time they spend on the scene is very variable. Two key points defining these time series are its beginning and its end, $[x_i(1),y_i(1)]$ and $[x_i(end),y_i(end)]$ as they define where the object is coming from and where it is going to. We formed a feature vector from this set of points and fed to a hierarchical clustering algorithm. For a data set made of m objects there are m*(m-1)/2 pairs in the dataset. We employed the Euclidean distance as a measure of similarity to calculate the distance between all object trajectories. Object trajectories with the minimum distance are clustered together. When two or more trajectories are set together its centroid is taken into account for further clustering. The successive merging of clusters is listed by the dendrogram. The evaluation of the dendrogram is typically subjective by adjudging which distance threshold appears to create the most natural grouping of the data. For this reason we have created an interface that allows the user to explore the dendrogram. The final number of clusters is set manually and typical values are between 12 to 25 for a data set of 1000 to 1500 mobile objects. To be noticed that as the aqcuisition performs in a multi-camera environment the clusters obtained can be generalised to different camera views thanks to a 3D calibration matrix applied during the on-line analysis system.

4.1.2 Contextual objects

A contextual object can be represented as a 12-tuple(c_id, c_type, c_start, c_end, c_involved_events_id, c_significant _event, c_rare_event, c_event_histogram, c_involved_

mobile_objects_id, c_histogram_mobile_objects, c_use_ duration, c_mean_time_of_use)
where
c_id; **c_type**; **c_involved_events_id**; **c_significant_event** are defined in the same way as for the mobile objects but referring to contextual objects.
The remaining fields indicate
c_start and **c_end** refer to the first and last instant the mobile object interacts with the contextual object
c_rare_event. This is the rarest event.
c_event_histogram. Gives the frequency of occurrence of all involved events.
c_involved_mobile_objects_id. All detected mobile objects interacting with the contextual object of interest.
c_histogram_mobile_objects. Gives the frequency of appearance for all involved mobile objects.
c_use_duration. Percentage of occupancy (or use of a contextual object). For instance, the Ticket Machine has a 10% of use over the observation time.
c_mean_time_of_use. Average time of interactions between the mobile object and the contextual object.

The contextual objects to be monitored are manually defined. This is a quick process and we also avoid computationally expensive algorithms. For the video sequences analysed in this work, the contextual objects are: 'Platform hall', 'Gates', 'VendingMachine1', and 'VendingMachine2'.

4.1.3 Video events

A video event can be represented as a 6-tuple(e_id, e_type, e_start, e_end, e_involved_mobile_object_id, e_involved _contextual_object_id)
where
e_id. The identifier label for the detected Event.
e_type. The class where the Event belongs to ('close_to', 'stays_at', …)
e_start. First moment on which the Event is detected.
e_end. Last moment on which the Event is seen.
e_involved_mobile_object_id. The identifier label of the object involved in that event.
e_involved_contextual_object_id. The name of the contextual object involved in that event.

4.2 Discovery of complex relationships

Once all statistical measures of the activities in the scene have been computed and the corresponding information is put into the proposed model format, we aim at discovering complex relationships that may exist between mobile objects themselves, and between mobile objects and contextual objects in the scene. For this task, the clustering methodology we decided to use is RARES (Relational Analysis And Regularized Similarity). This methodology gathers two different technologies: relational analysis theory and regularized similarity [3,4]. Relational analysis has been initiated and developed at the European Centre of Applied Mathematics (ECAM) at IBM France By F. Marcotorchino and P. Michaud in 1981 [14]. The principle of relational analysis consists in transforming the data usually represented as a $N \times M$ rectangular matrix where N is the number of objects and M is the number of variables measured on these objects to a $N \times N$ matrix representing a similarity measure for each pair of objects.
Each variable $V^k \ (k = 1,2,\ldots,M)$ is then transformed to a $N \times N$ matrix S^k where the term $s^k_{ii'}$ is the similarity measure between the two objects i and i' w.r.t. variable V^k. A dissimilarity measure $\bar{s}^k_{ii'}$ is then computed as the complement to the maximum similarity measure possible between these two objects. As the similarity between two different objects is less or equal to their self-similarities: that is $s^k_{ii'} \le \min(s^k_{ii}, s^k_{i'i'})$ then $\bar{s}^k_{ii'} = \min(s^k_{ii}, s^k_{i'i'}) - s^k_{ii'}$. The global similarity measure between objects i and i' over the M variable is $s_{ii'} = \sum_{k=1}^{M} s^k_{ii'}$ and their global dissimilarity is also $\bar{s}_{ii'} = \sum_{k=1}^{M} \bar{s}^k_{ii'}$. To cluster a population of N objects described by M variables, the relational analysis theory is based on the Condorcet criterion $C(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(s_{ii'}x_{ii'} + \bar{s}_{ii'}\bar{x}_{ii'}\right)$ where X is a binary $N \times N$ matrix representing the partition to discover in the data. The term $x_{ii'}$ is defined as follows:

$$x_{ii'} = \begin{cases} 1 & \textit{if } i \textit{ and } i' \textit{ are in the same cluster} \\ 0 & \textit{otherwise} \end{cases}$$

and $\bar{x}_{ii'} = 1 - x_{ii'}$
The mathematical formulation of the criterion to be maximised is:

$$\max(C(X)) \; w.r.t.$$
$$\begin{cases} x_{ii} = 1 & \textit{reflexivity} \\ x_{ii'} = x_{i'i} & \textit{symmetry} \\ x_{ii'} + x_{i'i''} - x_{ii''} \le 1 & \textit{transitivity} \end{cases}$$

Regularized similarity developed by H. Benhadda and F. Marcotorchino [3,4] is concerned with the weights of variables when one tries to perform a clustering task. Indeed, variables have intrinsic weights that are induced by their internal structures. For example if a categorical variable has 20 categories like the districts of Paris and another one has only two categories like the sexual gender, then it is more likely to meet randomly in Paris two persons of the same gender than two persons living in the same district. Thus a particular attention must be granted to these weights to not let some variables take more importance than what would be expected.

5 Results

5.1 On annotated data

We first tested the validity of our clustering algorithms (Hierarchical and Relational clustering) on labelled video

data. Caviar is an EC founded project that has made available a dataset of video clips with hand-labelled ground truth (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/). We focused our attention on the first part of the dataset, which contains people observed at the lobby entrance of a building. The annotated data include for each person its bounding box (id, centre coordinates, width, height, main axis orientation) with a description of his/her movement type (inactive, active, walking, running) for a given situation (moving, inactive, browsing) and with a given scenario context (browsing, immobile, left object, walking, drop down).

We applied first the hierarchical clustering algorithm to a dataset containing 104 persons or objects. We extracted their trajectories as the centre coordinates of the bounding box over time. We manually tuned the algorithm with the user interface to obtain 21 meaningful clusters. Figure 4 shows in the upper-left panel all trajectories from this dataset. The remaining plots in the figure are the three most common paths undertaken. As it can be observed clear trajectory patterns can be extracted from the clusters. In this case these three paths are characterised as cluster 8: 'entering right and up / exiting left', cluster 11: 'entering right bottom / exiting left', and cluster 15: 'entering right-middle / exiting right-bottom'.
We then applied relational analysis in order to obtain higher relations between objects. For this purpose we employed the object representation format described in 4.1.1 and because the annotated data is already available with a situation and context description, we generated the events such that each involved event is the concatenation of three pieces of information (movement's type, context and situation). This information is summarized in the table below:

Type	Context	Situation
(i)nactive	(b)rowsing	(m)oving
(a)ctive	(i)mmobile	(i)nactive
(w)alking	(l)eft object	(b)rowsing
(r)unning	(w)alking	
	(d)rop down	

Table 1: Semantic event information in CAVIAR.

For example, an Involved_Event having the value "awm" is related to an Active object in a Walking context and in a Moving situation. In the data we analysed, an object is involved in up to 12 such events during the observation time. A portion of the input data matrix is shown below in Table 2.

obj_id	obj_type	startframe	endframe	trajectory_type	obj_shape	Involved_Event1	Involved_Event2	Involved_Event3
84	person	13785	14312	19	big	wwm		
85	person	14338	14611	6	small	iii	aii	wim
86	person	14459	14523	16	tall	wwm		
88	person	14463	14684	7	big	rwm	wwm	
87	person	14491	14684	10	big	wim	aii	
89	person	14685	15097	15	tall	wim	iii	aii
90	person	14757	14938	16	tall	wwm		
93	person	15054	15542	16	big	wim	aii	wim
92	object	15256	15547	3	small	iii		
94	person	15548	15670	15	large	wim	aii	iii
95	person	15695	15750	14	small	iii	aii	iii
96	person	15817	16699	14	big	abb	wbm	abb
98	object	16051	16563	19	small	iii		
99	person	16920	17082	14	small	wim	aii	iii
103	person	16842	17513	4	small	wim	aii	iii
104	person	16966	17513	4	small	wim	aii	iii

Table 2: Input matrix used for the relational analysis clustering method. Remark that only some objects from the total detected set are represented in the table.

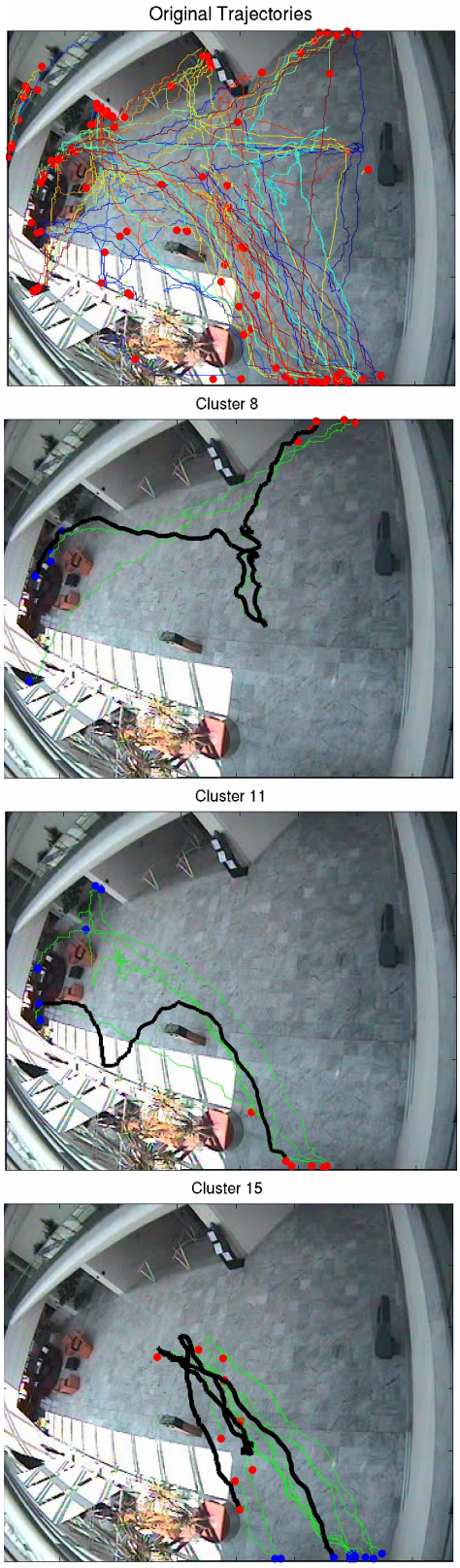
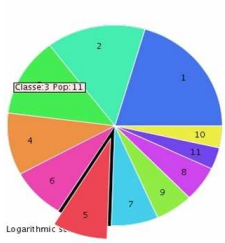


Figure 4. Original set of CAVIAR trajectories (upper-left panel) and three clusters showing most common undertaken paths.

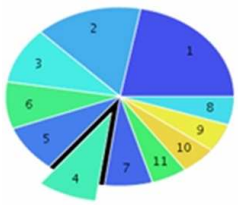


Typical Individual

Variable	Modality	Percent
STARTFRAME	11821-16920	100.0%
ENDFRAME	6653-19948	100.0%
S1	Inactive	100.0%
C1	Immobile	100.0%
T1	Inactive	100.0%
OBJ_TYPE	Object	100.0%
OBJ_SHAPE	Small	50.0%
TRAJECTORY_TYPE	3	25.0%

Figure 5: Resulting partition of the CAVIAR data after running the relational analysis algorithm. Properties of cluster 5 are given.

One of the clusters that RARES has discovered is presented in Figure 5. We can see that this cluster is made up of 4 items. All items are objects that were involved in only one Event and all of them were inactive, in an inactive situation and in an immobile context. This classification actually corresponds to the objects 'bag' that were annotated in the CAVIAR database.



Typical Individual

Variable	Modality	Percent
STARTFRAME	6654-11821	100.0%
ENDFRAME	6653-19948	100.0%
S2	Browsing	100.0%
C2	Walking	100.0%
T2	Running	100.0%
S1	Browsing	100.0%
C1	Walking	100.0%
T1	Walking	100.0%
OBJ_TYPE	Person	100.0%
TRAJECTORY_TYPE	8	80.0%
S3	Browsing	40.0%
C3	Walking	40.0%
T3	Walking	40.0%
OBJ_SHAPE	small	40.0%

Figure 6: Properties for cluster 4 in the CAVIAR data partition.

Two more clusters are shown below. In the first case, see Figure 6, all items are people that are involved in at least two events, first walking then running, in a browsing situation and within a walking context, 40 % of the people are involved in a third event of type walking with same situation and context. 80% of the total subjects had a trajectory with label number 8 'entering right and up-exiting left' shown in Figure 4.

In the second case, see Figure 7, at least 66% of the cluster members are involved in three events whose pattern type follows walking-inactive-walking and the same percentage of individuals are characterised by a trajectory with number label 15, 'entering right middle / exiting right bottom', also shown in Figure 4. We have thus gained the information that most persons with this trajectory type have a movement pattern walking-inactive-walking.

Typical Individual

Variable	Modality	Percent
S2	Inactive	100.0%
C2	Immobile	100.0%
C1	Immobile	100.0%
OBJ_TYPE	person	100.0%
S3	Browsing	83.0%
C3	Immobile	83.0%
T3	Walking	83.0%
S1	Browsing	83.0%
T1	Walking	83.0%
ENDFRAME	6653-19948	66.0%
T2	Inactive	66.0%
TRAJECTORY_TYPE	15	66.0%
OBJ_SHAPE	tall	50.0%
STARTFRAME	6654-11821	33.0%

Figure 7: Properties for cluster 3 in the CAVIAR data partition.

5.2 On large video recordings

In total we processed 73000 frames of video from the Torino underground (GTT, Italy), with an acquisition rate of 25 frames/s equals to about fifty minutes of video. We have analysed this period of time off-line. We applied our hierarchical clustering algorithm on the trajectories of mobile objects to obtain common behavioural paths undertaken by the people on the platform. Figure 8 presents the whole dataset of trajectories that we analysed. Using our interactive user interface we applied our hierarchical clustering selecting 22 clusters. The most common paths that people take are shown in Figure 9.

The clusters give clear semantic information on the behaviours undertaken by the metro users. For instance, Cluster 6 shown in upper panel of Figure 9 indicates that most users that buy a ticket, enter the station by the north doors. Cluster 3 indicates that after buying a ticket, users go straight to the gates to take the metro. Cluster 8 indicates that users

entering the station by the south doors go rather straight to the gates. Cluster 12 indicates that most users exiting the gates and leaving the station go through the south gates.

We further performed the statistical analysis of the interaction between users and the contextual objects. As mentioned in section 3.2 there are four contextual objects of interest in the scene, namely the Platform hall, the Gates, VendingMachine1 and VendingMachine2. Over the whole observation time a user was practically constantly present on the platform as we obtained a percentage of use of the platform of 91% of the observed time. The gates had a percentage of use of 36% indicating that the flow of people trough the gates was not constant over the observation time. The vending machines had a percentage of use of only 8% and 7% respectively indicating that most people did not stop long-time or did not stop at all to buy a ticket while in the station. This is further confirmed by the fact that most users buying tickets were detected as single subjects and came more rarely to the machines as groups. No crowd was directly detected at the vending machines however to a lower degree crowding was detected in the platform. No crowd was either detected at the gates.

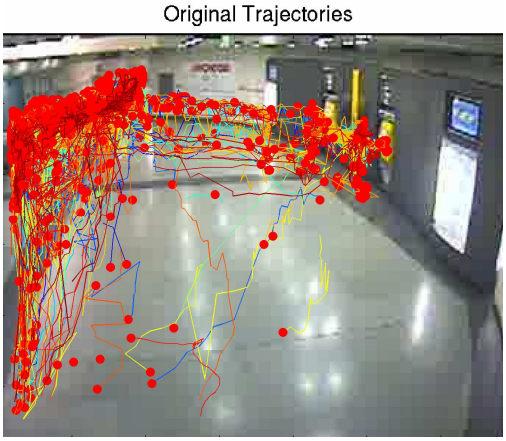


Figure 8: Original trajectories detected in one station of the Torino underground (1126 trajectories). Initial points of the trajectories are indicated by a point.

We did set up all knowledge in the format presented in section 4.1. A portion of the semantic tables obtained are presented next (Tables 3 to 5).

ctx_obj_id	ctx_obj_type	startframe	endframe	sig_evnt_type	rare_evnt_type
1	Platform	20050	92815	inside_zone(14486)	group_stays_inside_zone(30)
2	Gates	20055	92745	close_to(5103)	group_stays_at(40)
3	VendingMachine2	26560	90725	close_to(1020)	group_stays_at(200)
4	VendingMachine1	30680	90650	close_to(834)	group_stays_at(160)

ctx_obj_id	evnt_hist
1	inside_zone(14486) group_inside_zone(5523) crowd_inside_zone(1750) stays_inside_zone(1276) crowding_in_zone(109)
2	close_to(5103) stays_at(3489) group_close_to(329) group_stays_at(40)
3	close_to(1020) stays_at(490) group_close_to(341) group_stays_at(200)
4	close_to(834) stays_at(482) group_close_to(307) group_stays_at(160)

ctx_obj_id	mob_obj_hist	percnt_of_use	mean_time_of_use
1	Person(491) Unknown(102) PersonGroup(136) Luggage(34) Crowd(4)	91.1296%	35500.854
2	Person(135) Unknown(28) Luggage(12) PersonGroup(17)	36.8284%	35682.2396
3	Unknown(19) Person(15) Luggage(10) PersonGroup(2)	8.8704%	7401.087
4	Unknown(12) Person(11) Luggage(8)	7.7504%	16931.2903

Table 3: Contextual Objects semantic table.

Figure 9: Clusters showing the most common paths obtained from the dataset shown in figure 8. Initial trajectory points are indicated by coloured points. Final trajectory points are indicated by white points.

evnt_id	evnt_type	startframe	endframe	inv_objs_id	ctx_type
10161	inside_zone	39085	39085	1573	Platform
10162	close_to	39085	39085	1444	Gates
10163	inside_zone	39090	39090	1444	Platform
10164	inside_zone	39090	39090	1573	Platform
10165	close_to	39090	39090	1444	Gates
10166	inside_zone	39095	39095	1444	Platform
10167	inside_zone	39095	39095	1573	Platform
10168	close_to	39095	39095	1444	Gates
10169	inside_zone	39100	39100	1444	Platform
10170	inside_zone	39100	39100	1573	Platform
10171	close_to	39100	39100	1444	Gates
10172	group_inside	39105	39105	1573	Platform

Table 4: Events semantic table.

mob_obj_id	mob_obj_type	startframe	endframe	traj_type	shape_type	sig_evnt_id
6409	Person	81305	81365	2	small	inside_zone_Platform
6412	Person	81335	81335	1	tall	inside_zone_Platform
6392	PersonGroup	81130	81280	2	small	group_inside_zone_Platform
6400	PersonGroup	81210	81210	1	small	group_inside_zone_Platform
6399	Person	81200	81205	1	tall	inside_zone_Platform
6385	Person	81090	81150	2	tall	inside_zone_Platform
6381	PersonGroup	81030	81075	2	small	group_inside_zone_Platform
6374	Luggage	80950	81025	2	big	inside_zone_Platform
6370	Person	80900	80945	1	tall	inside_zone_Platform
6369	Luggage	80855	80935	2	large	inside_zone_Platform
6366	Person	80775	80775	1	tall	inside_zone_Platform
6353	Person	80205	80390	2	small	inside_zone_Platform
6361	Person	80310	80325	1	tall	inside_zone_Platform
5942	Person	76930	80265	2	tall	inside_zone_Platform

Table 5: Mobile Objects semantic table.

In the last step towards knowledge discovery we applied the relational analysis explained in section 4.2. The input matrix was the mobile objects semantic table presented above and containing in total 1126 detected objects. Some of the clusters returned are shown next (Tables 6 to 9).

Typical Individual

Variable	Modality	Percent
ENDFRAME	45845-84375	100.0%
STARTFRAME	48975-84220	100.0%
SIGN_EVENT_ID	inside_platform	92.0%
MOB_OBJ_TYPE	Person	83.0%
TRAJECTORY_TYPE	2	73.0%
SHAPE_TYPE	tall	28.0%

Table 6: The biggest cluster found (cluster 1) for the underground data that we input contains 287 persons. They were detected inside the platform and were associated to a trajectory cluster (shown in Figure 10) that indicated a user presence near the gates, which indicates the most frequent behaviour as most of the people detected in the station goes through the gates at some instant.

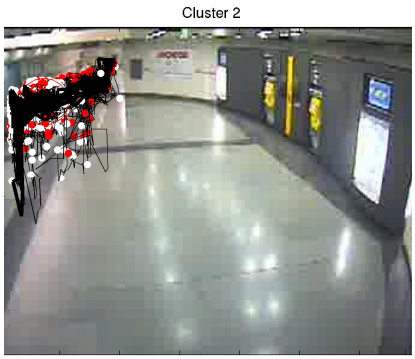


Figure 10: Trajectory type 2 'activity near the gates' associated to the biggest cluster found by the relational analysis algorithm.

Typical Individual

Variable	Modality	Percent
TRAJECTORY_TYPE	12	100.0%
SIGN_EVENT_ID	inside_platform	82.0%
MOB_OBJ_TYPE	person	60.0%
ENDFRAME	45845-84375	53.0%
STARTFRAME	48975-84220	50.0%
SHAPE_TYPE	small	32.0%

Table 7: Cluster 11 from the Torino data partition includes only 28 persons but they all have in common a trajectory of type 12 'exiting through south doors' (shown in Figure 9) and all persons were detected inside the platform.

Typical Individual

Variable	Modality	Percent
TRAJECTORY_TYPE	6	100.0%
ENDFRAME	45845-84375	60.0%
SIGN_EVENT_ID	inside_platform	60.0%
STARTFRAME	48975-84220	56.0%
MOB_OBJ_TYPE	person	46.0%
SHAPE_TYPE	small	30.0%

Table 8: Clusters 9 represents people in platform hall but associated with trajectories of type 6 'entering north doors – going to the vending (shown before in Figure 9).

Typical Individual

Variable	Modality	Percent
TRAJECTORY_TYPE	8	100.0%
ENDFRAME	45845-84375	64.0%
STARTFRAME	48975-84220	64.0%
SIGN_EVENT_ID	inside_platform	50.0%
MOB_OBJ_TYPE	Luggage	50.0%
SHAPE_TYPE	large	41.0%

Table 9: Cluster 8 represent people in platform hall but associated with trajectories of type 8 'entering south doors – going to the gates' (shown before in Figure 9).

Thus, this way the relational analysis can help us to group together people having similar behaviour. This is of particular interest to the end-user because significant events showing interactions with contextual objects are taken into account.

6 Conclusion

In this paper we have presented how knowledge discovery can be achieved on large recordings of video using an efficient knowledge representation format. The richness in the representation comes from the fact that both, moving objects and the contextual objects from the scene are studied together with their interaction. Yet, the proposed representation is clear as all activity knowledge is dissolved into three different semantic tables, namely mobile objects, contextual objects and video events. The proposed representation supports a rich set of spatial topological and temporal relations and captures not only quantitative properties but also higher semantic

concepts. Furthermore, a first layer of meaningful knowledge is directly extracted from the video streams and it already detects the interaction between moving objects and between contextual objects. A second layer of semantic knowledge is extracted by the off-line long term analysis of these interactions. First statistical information is obtained from the mobile objects and the contextual objects as well as their interactions. This is a major information source for the end-user. For instance, on large metro video recordings that we show, there is spatial and temporal information on the use of contextual objects. Mobile objects are also characterised by the trajectory they undertake which gives pertinent information about people behaviour and space occupancy. One hour of video takes approximately one minute to be treated by the hierarchical clustering algorithm in order to achieve the partition of the trajectory dataset. This is a reasonable processing time. We are currently analysing sequentially chunks of video of about one or two hours and then will further categorize the clusters in a temporal way for durations such as one day or one week. On a second step, the semantic knowledge gained from trajectory characterisation and statistical analysis can be used for the discovery of complex relationships. The relational analysis proposed in this paper shows to unlock hidden relations between people, their trajectories (behavioural information) and their significant associated interaction, between themselves or with contextual objects. Thus it can give a richer knowledge of the scene activity. A limitation of our system is due to the fact that the location of users is sometimes limited to a large portion of the scene like the platform. We will be looking to improve this point by finding a meaningful way to better partition those zones. We will also try to further characterise the user's trajectories in a similar way as Le et al. [10] in order to gain more information when applying the relational analysis algorithm. We also are planning to learn the different numerical values in our real time detection (T1, T2, …).

References

[1] A. H. R. Albers, R. G. J. Wijnhoven, E. G. T. Jaspers, P. H. N. de With. "Smart Search & Retrieval on Video Databases", *In Proc. of the IEEE International Conference on Consumer Electronics (ICCE)*, pp. 475-476, (2006).

[2] A. Avanzi, F. Bremond, M. Thonnat. "Tracking Multiple Individuals for Video Communication", *In Proc. of the IEEE International Conference on Image Processing*, **volume 2**, pp. 379-382, (2001).

[3] H. Benhadda. "La similarité régularisée et ses applications en classification automatique", PhD thesis of PARIS VI university (1998).

[4] H. Benhadda, F. Marcotorchino. "Introduction à la similarité régularisée en analyse relationnelle", Revue de Statistique Appliquée, **volume 46 (1)**, pp. 45-69, (1998).

[5] J.-F. Chen, H.-Y. Mark Liao; C.-W. Lin. "Fast Video Retrieval via the Statistics of Motion", *In Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, **volume 2**, pp. 437-440, (2005).

[6] H. D. Chon, D. Agrawal, A. El Abbadi. "Storage and retrieval of moving objects", *In Proc. of the International Conference on Mobile Data Management (MDM)*, pp. 173-184, (2001).

[7] F. Cupillard, F. Bremond, M. Thonnat. "Video understanding for metro surveillance". *In Proc of the IEEE International Conference on Networking, Sensing and Control, special session on Intelligent Transportation Systems (IC-NSC)*, **volume 1**, pp. 186-191 (2004).

[8] B. Georis, F. Bremond, M. Thonnat, B. Macq. "Use of an Evaluation and Diagnosis Method to Improve Tracking Performances", *In Proc. of the 3rd IASTED International Conference on Visualization, Imaging and Image Proceeding (VIIP)*, **volume 2**, (2003).

[9] G. Kollios, V. J. Tsotras, D. Gunopulos, A. Delis, M. Hadjieleftheriou. "Indexing animated objects using spatiotemporal access methods", *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, **volume 13 (5)**, pp. 758-777, (2001).

[10] T.-L. Le, A. Boucher, M. Thonnat. "Subtrajectory-based video indexing and retrieval", *In Proc. of Advances of Multimedia Modelling, 13th International Multimedia Modelling Conference*, **volume 1** , pp. 418-427, (2007).

[11] J. Z. Li, M. T. Ozsu, D. Szafron. "Modeling of moving objects in a video database", *In Proc. of the IEEE International Conference on Multimedia Computing and Systems (MMCS)*, pp. 336-343, (1997).

[12] C.-H. Lin, A. L. P. Chen. "Motion event derivation and query language for video databases", *Proceedings of SPIE, the International Society for Optical Engineering*, **volume 4315**, pp. 208-218, (2001).

[13] D. Liu, C. E. Hughes. "Deducing Behaviors from Primitive Movement Attributes", *Defense and Security Symposium, Proceedings of the SPIE*, **volume 5812**, pp. 180-189, (2005).

[14] F. Marcotorchino, P. Michaud. "Agrégation des similarités en classification automatique", Revue de Statistique Appliquée, **volume 30 (2)**, Dunod, (1981).

[15] D. Papadias, Y. Tao, J. Zhang, N. Mamoulis, Q. Shen, J. Sun. "Indexing and Retrieval of Historical Aggregate Information about Moving Objects", *IEEE Data Engineering Bulletin*, **volume 25 (2)**, pp. 10-17, (2002).

[16] J. Piater, S. Richetto, J. Crowley. "Event based activity analysis in live video using a generic object tracker". *In: Freyman, J. (ed.) Proc of the 3rd IEEE Workshop on performance evaluation of tracking and surveillance (PETS)*, (2002).

[17] Y. F. Tao, J. M. Sun, D. Papadias. "Analysis of predictive spatio-temporal queries", *ACM Transactions on database systems*, **volume 28 (4)**, pp. 295-336, (2003).

[18] S. A. Velastin, B. A. Boghossian, B. P. Lai Lo, J. Sun, M. A. Vicencio-Silva: PRISMATICA. "Toward Ambient Intelligence in Public Transport Environments", *IEEE Transactions Systems Man Cybernetics A*, **volume 35**, pp. 164-182, (2005).